Spring 2016

Using Penalized Regression to Uncover Peer Effects in the Spatial Autoregressive Model

Pawel Janas
pawel.janas@colorado.edu

Honor's Thesis in Economics

Using Penalized Regression to Uncover Peer Effects in the Spatial Autoregressive Model.

Pawel Janas[†]
Advised by Dr. Carlos Martins-Filho[*]

Committee: Dr. Will Kleiber[**]
Dr. Nicholas Flores[***]

Defense Date: 3/31/2016

Abstract

We propose a technique for estimating the spatial weights matrix (SWM) of the spatial autoregressive model (SAR) using the least absolute shrinkage and selection operator (Lasso), first proposed by Tibshirani (1996). The SWM is typically assumed *a priori* as a known matrix of correlations among spatially correlated data, as it cannot be estimated using standard techniques due to overfitting. However, we use the Lasso to discover the most prominent spatial coefficients in the SWM and estimate them using feasible generalized least squares, while setting the relatively unimportant effects to zero. Furthermore, the Lasso solutions are optimized to minimize mean squared error over a grid of possible spatial lag parameters. Finally, we use the LARS-Lasso algorithm to compute an illustrative example of the technique.

Keywords and phrases. Spatial autoregressive model, lasso, panel model, least angle regression, cross validation.

JEL classifications. C13, C31, C33.

Acknowledgments. I would like to to thank Dr. Martins-Filho for his unrelenting patience and guidance through this process. His input has been essential to this thesis. I would also like to thank Dr. Kleiber for introducing me to spatial statistics and Dr. Flores for organizing the honor's seminar where I was able to present my work. All errors and typos are mine.

[†] Department of Economics (BA), Department of Applied Mathematics (BS/MS), University of Colorado at Boulder
[*] Professor, Department of Economics, University of Colorado at Boulder
[**] Assistant Professor, Department of Applied Mathematics, University of Colorado at Boulder
[***] Professor and Department Chair, Department of Economics, University of Colorado at Boulder

1 Introduction

Obtaining large panels of time-series data is becoming easier and studying statistical models that glean information from such data is becoming more important to econometricians and statisticians. For example, researchers studying these data may be interested in the spatial-temporal dependence structure of the underlying data-generating process. Various models have been proposed to study the cross-sectional dependence of variables, including the spatial autoregressive model (SAR), see Elhorst [3].

Two important features of the SAR are the spatial weight matrix (SWM), which contains key information on how the response variables of the process are "connected" to one another, and the spatial lag parameter (ρ), which acts as a correlation multiplier. Typically, the first task in modeling a spatial process using the SAR is specifying the SWM using prior expert knowledge or imposing certain structures that are unique to process at hand. For example, a researcher studying physically spatial data may impose a contiguity structure: the (i, j) element of the SWM is set to one if i and j are neighbors and zero otherwise. Another specification is to populate the SWM with the inverse of a "distance" metric, such as Euclidean distance, between variables.

It is important to point out that "spatial" data includes much more that just data collected over a geographic region. Essentially, spatial data applies to stochastic processes that have no natural ordering, like time. Any set of variables that may exhibit correlation because of their "proximity" to one another can be viewed as spatial data. For example, students within the same school or firms within an industry may be considered spatially correlated variables regardless of where each student or firm is physically located. It may be valid to assume that students *within* a grade directly affect one another's academic performance: they form friend networks that influence their academic appetite. Likewise, students in different grades (e.g. a 1^{st} and 8^{th} grader) might not experience any pattern of correlation among each other. In section 2.1, I show that these effects lead to a sparse SWM.

In my thesis, I will explore how to *estimate* spatial effects from the panel SAR model through the SWM under a sparsity assumption. Unlike the current practice of imposing the SWM *a priori*, my goal is estimate it empirically using a penalized regression technique called the least absolute shrinkage and selection operator (Lasso), first proposed in Tibshirani [21]. The guiding research questions are: How do we estimate these peer effects while imposing a *minimal* amount about the structure on the spatial weight matrix of the SAR? Can we simultaneously estimate the spatial lag parameter and the SWM?

2 Model Specification

Autoregressive models are widely used in the time series analysis. Autoregression is a characteristic of a stochastic process in which the current state of a process depends on its value at some other state(s), e.g. $X_j = \alpha X_{j-1} + \epsilon_j$. Likewise, the SAR models autoregression in the spatial setting, where there is no natural ordering (like time) to the process. As such, the SAR allows each response variable of the process to affect any other response variables, which is the common setting in regional economics, network science, and the geosciences among many others. In effect, the SAR is a collection of simultaneously determined endogenous equations. The first equation of traditional SAR model can be written as:

$$Y_1 = \rho W_{12} Y_2 + \rho W_{13} Y_3 + ... + \epsilon_1 = \sum_{i=2}^{T} W_{1i} Y_i + \epsilon_1 \tag{1}$$

where W_{1i},...,W_{1T} are the autoregressive parameters $\in \mathbb{R}$, $\rho \in [-1, 1]$ is the spatial autocorrelation term, ϵ_1 is the unobservable disturbance term and T is the number of random variables under consideration.

Stacking the Ys, the traditional SAR model can be written in matrix notation as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix} = \rho \begin{bmatrix} 0 & W_{12} & \cdots & W_{1T} \\ W_{21} & 0 & \cdots & W_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ W_{T1} & W_{T2} & \cdots & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix} \tag{2}$$

2.1 Panel SAR Model

The model that is of interest in this proposal is the SAR model for panel data, or panel SAR (PSAR). Panel data is simply cross-sectional time series data: observations from multiple processes are collected through time. An example of panel data is the collection of observations (such as test scores) from each entering class (panel) within a school, from the first year until graduation. In theory, both intra-panel spatial correlation (students within a grade affect one another) and inter-panel spatial correlation (students affect one another through grades) may both be present.

Let j denote a panel and i an observation within a panel, and let the doubly indexed data be Y_{ij}. Let T_j denote the size of panel j. For notation purposes, define Y_j and ϵ_j as following: $Y_j = \begin{bmatrix} Y_{1j} & Y_{2j} & \cdots & Y_{Tj} \end{bmatrix}^T$ and $\epsilon_j = \begin{bmatrix} \epsilon_1^j & \epsilon_2^j & \cdots & \epsilon_{Tj}^j \end{bmatrix}'$. To illustrate, the 1^{st} equation of panel j

of the full PSAR looks like the following:

$$Y_{1j} = \rho_j W_{12}^j Y_{2j} + \rho_j W_{13}^j Y_{3j} + ... + \rho_j W_{1T_j}^j Y_{Tj} + \epsilon_1^j = \rho_j \begin{bmatrix} 0 & W_{12}^j ... & W_{1T}^j \end{bmatrix} \begin{bmatrix} Y_{1j} & Y_{2j} & ... & Y_{Tj} \end{bmatrix}' + \epsilon_1^j \tag{3}$$

where ρ_j is the spatial correlation parameter of panel j. Then, the 1^{st} panel SAR is:

$$Y_1 = \rho_1 \underbrace{\begin{bmatrix} 0 & W_{12}^1 & \cdots & W_{1T}^1 \\ W_{21}^1 & 0 & \cdots & W_{2T}^1 \\ \vdots & \vdots & \ddots & \vdots \\ W_{T1}^1 & W_{T2}^1 & \cdots & 0 \end{bmatrix}}_{X_1} \cdot Y_1 + \underbrace{\begin{bmatrix} \epsilon_1^1 \\ \epsilon_2^1 \\ \vdots \\ \epsilon_{T_1}^1 \end{bmatrix}}_{\epsilon_1} \tag{4}$$

Now, let $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}$. The full PSAR model becomes:

$$\mathbf{Y} = \underbrace{\begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_N \end{bmatrix}}_{\mathbf{X}} \mathbf{Y} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}}_{\Psi} \tag{5}$$

To makes the exposition of the rest of the paper more clear, we set $T_j = T \ \forall j$, though this assumption is not necessary for the results that follow.

Assumption 1.

$$X_j = X_{j'} \qquad \forall j \neq j'$$

This assumption states that the affect of one student on another student is constant through the panels (grades). We drop the superscript on W and group individual 1 in a vector to write (5) as:

$$\tilde{Y}_1 = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1N} \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_1 Y_{21} & \rho_1 Y_{31} & \cdots & \rho_1 Y_{T1} \\ \rho_2 Y_{22} & \rho_2 Y_{32} & \cdots & \rho_2 Y_{T2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_N Y_{2N} & \rho_N Y_{3N} & \cdots & \rho_N Y_{TN} \end{bmatrix}}_{X_1(\rho)} \underbrace{\begin{bmatrix} W_{12} \\ W_{13} \\ \vdots \\ W_{1T} \end{bmatrix}}_{\beta_1} + \underbrace{\begin{bmatrix} \epsilon_1^1 \\ \epsilon_1^2 \\ \vdots \\ \epsilon_1^N \end{bmatrix}}_{\epsilon_1} \tag{6}$$

with $\tilde{\mathrm{Y}}_1 \in \mathbb{R}^{\mathrm{N}}$, $X_1 \in \mathbb{R}^{\mathrm{N\ x\ T\text{-}1}}$, $\beta_1 \in \mathbb{R}^{\mathrm{T\text{-}1}}$ and $\epsilon_1 \in \mathbb{R}^{\mathrm{N}}$. Note that $X_i(\rho)$ is constructed such that i^{th} individual's observations are dropped. That is, the model for individual 2 may be written as:

$$\tilde{\mathrm{Y}}_2 = \begin{bmatrix} Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2N} \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_1 Y_{11} & \rho_1 Y_{31} & \cdots & \rho_1 Y_{T1} \\ \rho_2 Y_{12} & \rho_2 Y_{32} & \cdots & \rho_2 Y_{T2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_N Y_{1N} & \rho_N Y_{3N} & \cdots & \rho_N Y_{TN} \end{bmatrix}}_{X_2(\rho)} \underbrace{\begin{bmatrix} W_{21} \\ W_{23} \\ \vdots \\ W_{2T} \end{bmatrix}}_{\beta_2} + \underbrace{\begin{bmatrix} \epsilon_2^1 \\ \epsilon_2^2 \\ \vdots \\ \epsilon_2^N \end{bmatrix}}_{\epsilon_2} \tag{7}$$

Stacking $[\tilde{\mathrm{Y}}_1, \tilde{\mathrm{Y}}_2,...,\tilde{\mathrm{Y}}_T]^T$, we get the following:

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathrm{Y}}_1 \\ \tilde{\mathrm{Y}}_2 \\ \vdots \\ \tilde{\mathrm{Y}}_T \end{bmatrix} = \underbrace{\begin{bmatrix} X_1\rho & 0 & \cdots & 0 \\ 0 & X_2(\rho) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_T(\rho) \end{bmatrix}}_{\mathbf{X}(\rho)} \cdot \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix}}_{U} \tag{8}$$

with $\tilde{\mathbf{Y}} \in \mathbb{R}^{\mathrm{NT}}$, $\mathbf{X}(\rho) \in \mathbb{R}^{\mathrm{NT\ x\ T(T\text{-}1)}}$, $\beta \in \mathbb{R}^{\mathrm{T(T\text{-}1)}}$, and $\Omega \in \mathbb{R}^{\mathrm{NT}}$. Note that we can arrange β in such a way to fill the SWM in (2). In fact, SWM $= \begin{bmatrix} \beta_1' \\ \beta_2' \\ \vdots \\ \beta_T' \end{bmatrix}$.

Assumption 2.

For finite N and T, we require:

$$(1) \qquad |\beta|_c \leqslant NT$$

where $|\beta|_c$ denotes the cardinality of β, i.e. the number of non-zero elements. Notice that (1) is the standard rank identification assumption, as we cannot estimate β in (8) uniquely unless $NT \geqslant T(T-1)$, which implies, $N \geqslant T-1$.

Assumption 3.

$$\mathbf{E}[\epsilon_j] = 0 \qquad \mathbf{E}[\epsilon_i \epsilon_i^T] = \sigma_{ii} \cdot \mathbb{I}_N \quad \mathbf{E}[\epsilon_i \epsilon_j^T] = \sigma_{ij} \cdot \mathbb{I}_N \ \ \forall i \neq j$$

Equivalently, let $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1T} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{T1} & \sigma_{T2} & \cdots & \sigma_{TT} \end{bmatrix}$. Then, $\mathbf{E}[\Omega\Omega^T] = \Sigma \otimes \mathbb{I}_N$

Assumption 4. The matrix $\mathbf{X}(\rho)$ is of full rank.

2.2 Estimation of the model via the Lasso

Estimation of (8) requires estimates for β, Σ, and ρ. We propose the following steps:

1. For a fixed choice of ρ, compute the Lasso estimate for each β_i for $i = 1, 2, ..., T$.

$$\widehat{\beta}_i = \min_{\beta_i \in \mathbb{R}^T} \{ \frac{1}{2} \| \tilde{\mathrm{Y}}_1 - X_i(\rho)\beta_i \|_2^2 + \lambda \|\beta_i\|_1 \tag{9}$$

2. Update X_i to include only covariates of non-zero coefficients post lasso. Store the indeces of zero coefficients. Compute the method of moments estimator for Σ as:

$$\widehat{\sigma_{ij}} = \frac{1}{N} \left(\tilde{\mathrm{Y}}_i - X_i(\rho)\widehat{\beta}_i \right)^T \left(\tilde{\mathrm{Y}}_j - X_j(\rho)\widehat{\beta}_j \right)$$

3. Re-compute $\widehat{\beta}$ for post-Lasso coefficients using the generalized least squares estimator. Note that $\mathbf{X}(\rho)$ includes only covariates of non-zero coefficients post Lasso:

$$\widehat{\beta}_{FGLS} = (\mathbf{X}(\rho)^T (\widehat{\Sigma}^{-1} \otimes \mathbb{I}_N) \mathbf{X}(\rho))^{-1} (\mathbf{X}^T (\widehat{\Sigma}^{-1} \otimes \mathbb{I}_N) \tilde{\mathbf{Y}})$$

   Combine the zero coefficients and $\widehat{\beta}_{FGLS}$ in the original order to form $\widehat{\beta}$.

4. Repeat 1-3 for all ρ on some predefined grid, such as [{-0.5}, {0.5}].

5. Select $(\widehat{\rho}, \widehat{\beta})$ that minimize squared error, i.e. $\left(\tilde{\mathbf{Y}} - \mathbf{X}(\widehat{\rho})\widehat{\beta} \right)^T \left(\tilde{\mathbf{Y}} - \mathbf{X}(\widehat{\rho})\widehat{\beta} \right)$

3 The Lasso

One goal of regression is to predict. Assume Y = f(X) + ϵ where f is a function of the data and E[ϵ] = 0 with Var(ϵ) = σ^2. Prediction error can be defined as:

$$Err(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0]$$

$$Err(x_0) = \sigma^2 + Bias^2(\hat{f}(x_o)) + Var(\hat{f}(x_0))$$

In the presence of a large number of covariates, ordinary least squares (OLS) regression has two main shortcomings when it comes to prediction: (1) generally high variance (but low finite sample bias) and (2) lack of interpretable results, especially when the coefficient estimates are close to zero. Two methods have been developed to deal with these issues individually. Ridge regression shrinks the coefficients towards zero (but does not set them to zero) by imposing an L_2 penalty on OLS, thereby reducing variance at the cost of bias. More, methods of subset selection fit sets of regressors to the model consecutively. These methods are unstable and computationally infeasible with a large number of possible regressors.

Tibshirani [21] proposes a new method that combines the two methods described above into a stable subset selection process called the least absolute shrinkage and selection operator (Lasso), defined as the OLS estimate with an L_1 norm penalty on the coefficients:

$$\begin{aligned} &\{1\} \quad \widehat{\beta}_1 = \arg\min_{\beta}\{\|\tilde{\mathbf{Y}} - \mathbf{X}(\rho)\beta\|_2^2 + \lambda\|\beta\|_1\} \quad \text{or} \\ &\{2\} \quad \widehat{\beta}_2 = \arg\min_{\beta}\{\sum_{i=1}^{NT}(\tilde{\mathbf{Y}}_i - \sum_j \beta_j \mathbf{X}(\rho)_{ij})^2\} \qquad \text{subject to} \quad \sum_j |\beta_j| \leqslant t \end{aligned} \tag{10}$$

where $\lambda, t \geqslant 0$ control the amount of "shrinkage" and are chosen through cross-validation. I refer the reader to section 8 for more information about the Lasso.

4 Computing the Lasso

4.1 Feasibility of Descent Algorithms for the Lasso - KKT Conditions

Many convex optimization problems can be solved using descent methods, which are essentially search algorithms with stopping criteria based on the Karush-Kuhn-Tucker conditions. Recall our Lasso problem for individual i. For notational clarity, drop the index i:

$$\widehat{\beta} = \min_{\beta \in \mathbb{R}^{\mathrm{T}}}\{\frac{1}{2}\|\tilde{\mathrm{Y}} - X\beta\|_2^2 + \lambda\|\beta\|_1\} \tag{11}$$

which is equivalent to (see section 8 for proof):

$$\widehat{\beta} = \min_{\beta \in \mathbb{R}^{\mathrm{T}}}\{\frac{1}{2}\|\tilde{\mathrm{Y}} - X\beta\|_2^2\} \quad \text{subject to :} \quad \|\beta\|_1 \leqslant t \tag{12}$$

Remark. KKT Conditions that satisfy (11) are:

$$X_j^T(\tilde{\mathrm{Y}} - X\beta) = \lambda u_j \qquad j = 1, ..., T \tag{13}$$

where X_j denotes the jth column of X and u is the differential of $\|\beta\|_1$:

$$u_j = \begin{cases} sign(\beta_j) & \text{if } \beta_j \neq 0 \\ z \in \mathbf{R} : |z| \leqslant 1 & \text{if } \beta_j = 0 \end{cases} \tag{14}$$

Proof. In general, the optimization with M constraints:

$$\min f(x) \qquad \text{subject to} \qquad h_k(x) \leqslant 0 \quad k = 1, ..., M$$

has the following KKT conditions. which involve subdifferentials as L_1 is not differentiable at 0:

- (stationarity) $0 \in \partial f(x) + \sum_{i=1}^{m} \lambda_i \partial h_i(x)$
- (complementary slackness) $\lambda_i * h_i(x) = 0 \qquad \forall i$
- (dual feasibility) $\lambda_i \geqslant 0 \qquad \forall i$

In our case, let $f(\beta) = \frac{1}{2}\|\tilde{\mathrm{Y}} - X\beta\|_2^2$. Then,

$$f(\beta) = \frac{1}{2}(\tilde{\mathrm{Y}} - X\beta)^T(\tilde{\mathrm{Y}} - X\beta) = \tilde{\mathrm{Y}}^T\tilde{\mathrm{Y}} - \tilde{\mathrm{Y}}^T X\beta - \beta^T X^T\tilde{\mathrm{Y}} + \beta^T X^T X\beta \tag{15}$$

$$\frac{\partial f(\beta)}{\beta} = \frac{1}{2}(-2X^T\tilde{\mathrm{Y}} + 2X^T X\beta) = X^T(\tilde{\mathrm{Y}} - X\beta) \tag{16}$$

$$\frac{\partial f(\beta)}{\beta_j} = X_j^T(\tilde{\mathrm{Y}} - X\beta) \tag{17}$$

Now let $h(\beta) = |\beta|_1$.

$$h(\beta) = |\beta_1| + |\beta_2| + ... + |\beta_T| \tag{18}$$

$$\frac{\partial h(\beta)}{\partial \beta_i} = \frac{\partial}{\partial \beta_i}|\beta_i| = \begin{cases} sign(\beta_j) & \text{if } \beta_j \neq 0 \\ z \in \mathbf{R} : |z| \leqslant 1 & \text{if } \beta_j = 0 \end{cases} \tag{19}$$

Note that the Kuhn Tucker conditions imply 2^T constrains on β, as sign(β) can take on two values for each β_j, $j = 1, 2, ..., T$. For large T, doing subset selection (i.e. testing all combinations of β_j that minimize some criterion) would be computationally infeasible.

□

4.2 Least Angle Regression Lasso (LARS) Algorithm

Though standard quadratic programming techniques can be used to solve for the Lasso, the LARS, developed by Efron et al [2], is a computationally faster algorithm that can be trivially adapted to the Lasso context.

First, we assume that X and each $\tilde{\mathrm{Y}}_i$ have been standardized to have mean 0 and unit length through location and scale transformations:

$$\sum_{i=1}^{T} \tilde{\mathrm{Y}}_j(i) = 0 \qquad \sum_{i=1}^{N} X(\rho)_{ij} = 0 \qquad \sum_{i=1}^{N} X(\rho)_{ij}^2 = 1 \quad \forall j = 1, ..., T$$

A maximum of T steps are required for each individual's Lasso computation.

Let $\hat{u} = X(\hat{\rho})\hat{\beta}$ and $C(\hat{u}_k) = X(\hat{\rho})'(\tilde{\mathrm{Y}}_i - \hat{u}_k)$ be the vector of correlations between the residual at step k and all available covariates. Let $u_0 = 0$ and $\beta_L = 0$. The LARS procedure, for a given λ, works as follows:

1. Find the covariate most correlated with $\tilde{\mathrm{Y}}$, i.e. find $\max(C(\hat{u}_0))$, say $x_1 = X(\rho)_{i1}$
2. Go in the direction of x_1, i.e. $\hat{u}_1 = \hat{u}_0 + \hat{\beta}_1 x_1$. Recompute the new residual. Increment $\hat{\beta}_1$ until $C(\hat{u}_0) = C(\hat{u}_1)$, that is, until some other covariate has as much correlation with the new residual as did the previous predictor with the previous residual.
3. Repeat step 2 until a third covariate has as much correlation with the residual formed by the joint direction of x_1 and x_2 as did the previous estimate. Keep repeating step 2 as long as $\sum_{i=1}^{k} \hat{\beta}_i \leqslant \lambda$.
4. If any coefficient $\hat{\beta}_i$ during a run hits zero, drop this covariate from the set of available covariates. This is the Lasso correction for the Lasso-LARS algorithm. For the last iteration, set $\hat{\beta}_p$ such that $\hat{u}_p$ equals the projection of y onto the subspace spanned by the covariates chosen up to that point.
5. $\hat{\beta}_L = \langle \hat{\beta}_1, ..., \hat{\beta}_p \rangle$

We optimally choose λ through cross-validation.

4.3 Cross-validation

Here I describe how the tuning parameter in equation (1), t, is chosen in practice. This process is called cross-validation, which is covered in Hastie et al. [7]

1. Discretize t on a reasonable set, such as t ∈ (0,10)

2. Divide the samples into K folds (groups) of roughly equal size at random.

3. For each fold k = 1, 2, . . . ,K and for each t

   a) Estimate the Lasso using the data in all folds except for fold k

   b) Use the estimated Lasso from (a) to predict in fold k and calculate the total error in this fold:

$$\epsilon_k(\theta) = \sum_{i \in k} (y_i - X\beta_L(t))^2$$

4. Average the total error over all folds, then increment in t. The cross-validated value of t will be the one for which the Lasso estimate produces the smallest average total error.

4.4 Pseudocode

Algorithm 1 Model Estimation via Lasso

procedure
  Specify ρ set, e.g. [-0.5, 0.5]
  $\rho \leftarrow$ All combinations of ρ set

  for k = 1 to size(ρ) **do**
    Compute $\mathbf{X(\rho)} \leftarrow$ (8) with $\rho = \rho(k)$

    for i = 1 to T **do**
      $X_i \leftarrow X$ with ith column removed
      $Y \leftarrow \tilde{\mathrm{Y}}_i$
      $\widehat{\beta}_i \leftarrow$ LARS-Lasso(X_i,Y), as described above, with 5-fold cross-validation
      Update X_i to include only covariates of non-zero coefficients post-lasso

    for i,j = 1 to T **do**
      $\sigma_{ij} \leftarrow \frac{1}{N}\left(\tilde{\mathrm{Y}}_i - X_i(\rho)\widehat{\beta}_i\right)^T \left(\tilde{\mathrm{Y}}_j - X_j(\rho))\widehat{\beta}_j\right)$

    Compute FGLS of β: $\widehat{\beta} \leftarrow (\mathbf{X(\rho)}^T(\widehat{\Sigma}^{-1} \otimes \mathbb{I}_N)\mathbf{X(\rho)})^{-1}(\mathbf{X(\rho)}^T(\widehat{\Sigma}^{-1} \otimes \mathbb{I}_N)\tilde{\mathbf{Y}})$
    Compute and store MSE: $MSE(\widehat{\beta}, \widehat{\rho}) \leftarrow \left(\tilde{\mathbf{Y}} - \mathbf{X}(\rho)\widehat{\beta}\right)^T \left(\tilde{\mathbf{Y}} - \mathbf{X}(\rho)\widehat{\beta}\right)$
  return $(\widehat{\beta}, \widehat{\rho}) = \min MSE(\widehat{\beta}, \widehat{\rho})$

5 Example and Discussion

We generated NT = 120 spatially dependent data using the following ρ and randomly generated, sparse SWM:

Table 1: True SWM and ρ

N = 12	ρ	0.6	0	0.2	-0.5	0.3	0.7	0	0	0.9	-0.9	0.4	0
T = 10	β_1	0	-0.32	0.21	0	0	0	0.60	0	-1.11	-0.22		
	β_2	0.11	0	-1.66	0.08	-0.67	0.23	0	-0.75	-0.15	0		
	β_3	0.79	0.31	0	0.78	-0.68	0	0	-0.9	0	-1.47		
	β_4	0.33	0.22	0	0	-0.09	0	-0.15	0	0	0.16		
	β_5	0	-0.35	-0.25	-0.49	0	0	-0.33	-0.54	0	0		
	β_6	0	2.71	0	0	0.63	0	0.66	1.50	0	0		
	β_7	0	0	-0.21	0.69	-0.06	0	0	0.12	0	0		
	β_8	1.31	0	0	-1.25	-1.04	1.97	0	0	-0.63	0		
	β_9	1.93	-1.51	0	0	-0.1726	1.24	0	-0.53	0	-0.75		
	β_{10}	0.52	1.28	-0.26	-0.60	0	0	0	0	-0.11	0		

with $\epsilon_i \sim$ i.i.d. $N(1, 0.01)$ for all i. Specifically, we generated ϵ and computed Y using a re-arranged version of (5):

$$Y = (\rho \otimes \mathbb{I}_T)\mathbf{W}\mathbf{Y} + \Psi \qquad (\mathbb{I} - (D_\rho \otimes \mathbb{I}_T)\mathbf{W})\mathbf{Y} = \Psi \qquad \mathbf{Y} = (\mathbb{I} - (D_\rho \otimes \mathbb{I}_T)\mathbf{W})^{-1}\Psi$$

$$\mathbf{W} = \begin{bmatrix} W & 0 & \cdots & 0 \\ 0 & W & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & W \end{bmatrix} \qquad W = \begin{bmatrix} 0 & W_{12} & \cdots & W_{1T} \\ W_{21} & 0 & \cdots & W_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ W_{T1} & W_{T2} & \cdots & 0 \end{bmatrix}$$

Keeping in mind that each iteration of Algorithm 1 takes about 3.5 seconds to run on the computer available to us (MacBook Pro with a 2.4 GHz Intel Core 2 Duo processor and 4 GB of memory), we restricted our ρ set to [-0.5, 0.5]. This produced $2^{12} = 4096$ runs and Algorithm 1 took around 4 hours to compute. (For comparison, having three elements in the ρ set would require 22 days of computation.)

5.1 Estimated SWM

In general, the algorithm produced estimators of SWM that were more sparse than the true SWM. The best estimator is shown in Table 2. It correctly identified 69% of the zero coefficients and 46% of the non-zero coefficients. However, it also mis-identified 58% of the zero entries (incorrectly

estimated as zeros/true number of zeros) and 28% of the non-zero entries (incorrectly estimated as non-zeros/true number of non-zeros). The cross-validated λs are given in Figure 4. The green line depicts the lambda that produced the smallest average MSE over a 5-fold cross validation and the blue line depicts the largest lambda within one standard error of the MSE of the green.

Table 2: Estimated SWM

N = 12	$\widehat{\rho}$	0.5	-0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.5	-0.5
T = 10	$\widehat{\beta_1}$	0	-0.27	0	0	0	0	0	0	-0.51	0.02		
	$\widehat{\beta_2}$	0	0	0	0	0	0	0	0	0	0		
	$\widehat{\beta_3}$	2.16	-3.56	0	-0.26	0	0.48	-1.54	-0.31	1.65	1.57		
	$\widehat{\beta_4}$	0	0	0	0	0	0	0	0	0	0		
	$\widehat{\beta_5}$	0	0	0	0	0	0	0	0	0	0		
	$\widehat{\beta_6}$	0	-0.46	0	0	0	0	0	0	0	-0.01		
	$\widehat{\beta_7}$	0	0	-0.10	0	-0.18	0	0	0	0	0		
	$\widehat{\beta_8}$	-0.94	0.30	-1.86	0.66	1.16	-0.47	0.81	0	0	-0.85		
	$\widehat{\beta_9}$	1.85	0	0.93	-1.32	0.91	0.79	-2.96	0	0	3.31		
	$\widehat{\beta_{10}}$	0	0	0.11	0	0	0	0	-0.17	0	0		

For a visual comparison, we produced Table 3. The green entries correspond to the correct estimated coefficients, in terms of sign, without regard to magnitude. The right-most column totals the number of "correctly" identified coefficients. It is clear that this estimator does not estimate the true SWM very well. We believe there were three main issues: (1) extremely finite data size (120 parameters for 90 possible parameters), (2) the ρ set was not sufficiently large enough due to computational limitations and, lastly, (3) the Lasso-LARS algorithm performs sub-optimally with linearly dependent covariates [2]. Issue (1) is evident in Figure 4, with large standard errors on the cross-validated λ. Each fold in the cross-validation process contained fewer than five observations, which led to imprecise specification. Both issues (2) and (3) can be mitigated in future work. In some scenarios, researchers may not be interested in ρ and disregard its estimation. In terms of computation, newer algorithms have recently been proposed that outperform the Lasso-LARS, such as coordinate descent methods (see for example ([15])).

5.2 Analysis of MSEs and Groupings

We plotted a histogram of the MSEs for each possible ρ (Figure 1) and noticed that a small number of combinations produced very similar MSEs in the range (0.18-0.24). A natural question arose: is there an estimator that better identifies the SWM with a slightly higher MSE? Unfortunately, as depicted in Figure 3 and Tables 4/5, the estimators with MSEs in the range (0.18-0.24) performed similarly. In addition, as Lasso solutions become more sparse than the truth, we expect

the MSEs to rise. This phenomenon is depicted in Figure 2. As the proportion of correctly and incorrectly identified zero coefficients tends to 1, the MSE of these solutions increases.

Table 3: Estimated SWM, Visual Comparison

N = 12	$\widehat{\rho}$	+	-	+	+	+	+	+	+	+	-	+	-
T = 10	$\widehat{\beta_1}$	0	-0.27	0	0	0	0	0	0	-0.51	0.02		7
	$\widehat{\beta_2}$	0	0	0	0	0	0	0	0	0	0		3
	$\widehat{\beta_3}$	2.16	-3.56	0	-0.26	0	0.48	-1.54	-0.31	1.65	1.57		3
	$\widehat{\beta_4}$	0	0	0	0	0	0	0	0	0	0		5
	$\widehat{\beta_5}$	0	0	0	0	0	0	0	0	0	0		5
	$\widehat{\beta_6}$	0	-0.46	0	0	0	0	0	0	0	-0.01		5
	$\widehat{\beta_7}$	0	0	-0.10	0	-0.18	0	0	0	0	0		8
	$\widehat{\beta_8}$	-0.94	0.30	-1.86	0.66	1.16	-0.47	0.81	0	0	-0.85		1
	$\widehat{\beta_9}$	1.85	0	0.93	-1.32	0.91	0.79	-2.96	0	0	3.31		3
	$\widehat{\beta_{10}}$	0	0	0.11	0	0	0	0	-0.17	0	0		4

Table 4: Results for 10 of the 20 estimators with smallest MSE

MSE	0.23	0.24	0.20	0.22	0.19	0.21	0.22	0.23	0.23	0.21
% Correct Zero	70.83	70.83	79.17	70.83	70.83	70.83	72.92	72.92	72.92	72.92
% Correct Non-Zero	32.69	32.69	38.46	32.69	32.69	32.69	32.69	32.69	32.69	32.69
% Incorrect Zero	72.92	72.92	66.67	72.92	72.92	72.92	72.92	72.92	72.92	72.92
% Incorrect Non-Zero	26.92	26.92	19.23	26.92	26.92	26.92	25.00	25.00	25.00	25.00

Table 5: Results for the remaining 10 of the 20 estimators with small MSE, smallest in bold.

MSE	0.20	0.22	0.20	0.23	0.20	**0.18**	0.20	0.24	0.24	0.24
% Correct Zero	77.08	68.75	68.75	68.75	68.75	**68.75**	68.75	68.75	68.75	68.75
% Correct Non-Zero	26.92	46.15	46.15	46.15	46.15	**46.15**	46.15	46.15	46.15	46.15
% Incorrect Zero	79.17	58.33	58.33	58.33	58.33	**58.33**	58.33	58.33	58.33	58.33
% Incorrect Non-Zero	21.15	28.85	28.85	28.85	28.85	**28.85**	28.85	28.85	28.85	28.85

6 Conclusion

The motivation behind this work was to propose an estimation procedure to estimate the spatial weights matrix and the spatial parameters, which are fixtures in econometric spatial autoregressive models. Instead of assigning the values in the SWM using "expert knowledge" or contiguity measures, we sought to these estimate spatial effects purely from the spatial data. Using panel data and assuming that the true SWM is sparse, we proposed a multi-step procedure using the Lasso. First, the Lasso was used to eliminate the non-important covariates on the "individual" level. We then used the Lasso solutions to estimate the covariance structure of the disturbances and using this estimate to compute the GLS estimate of the coefficients. Finally, we conducted a test to see if our estimation is reasonable, computing the Lasso using the LARS algorithm. The estimation was not perfect and we were limited by two main factors: (1) constrained size of the grid of possible ρ, (2) the strong dependence of the covariates.

7 Figures

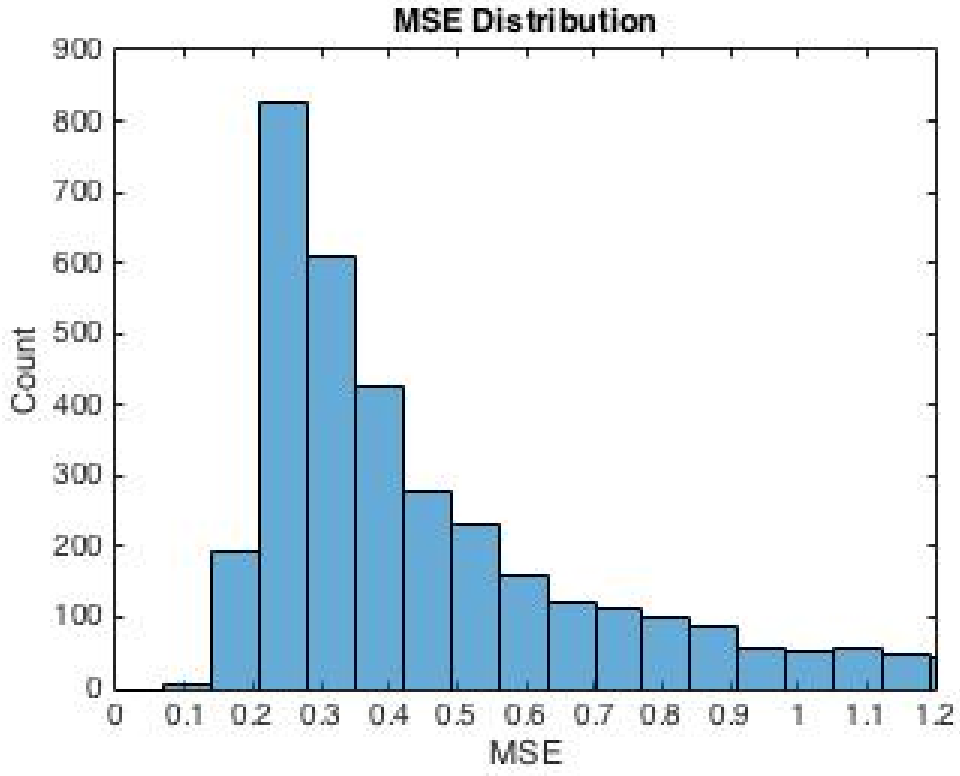


Figure 1: Distribution of MSEs for n = 4096 rho combinations

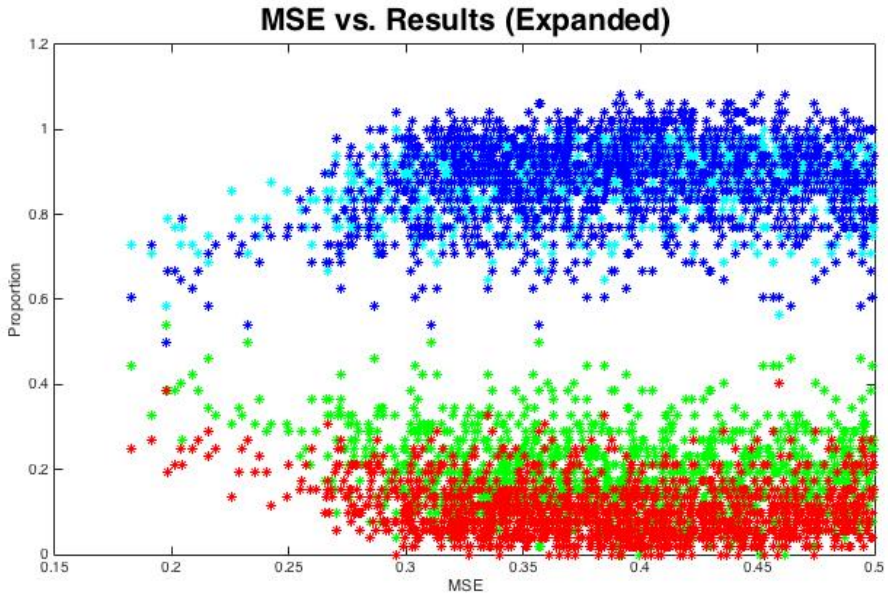


Figure 2: Results vs. MSE for MSE $\leqslant 0.5$.
Cyan - % Correct Zero, Green - % Correct Non-Zero, Blue - % Incorrect Zero, Red - % Incorrect Non-Zero

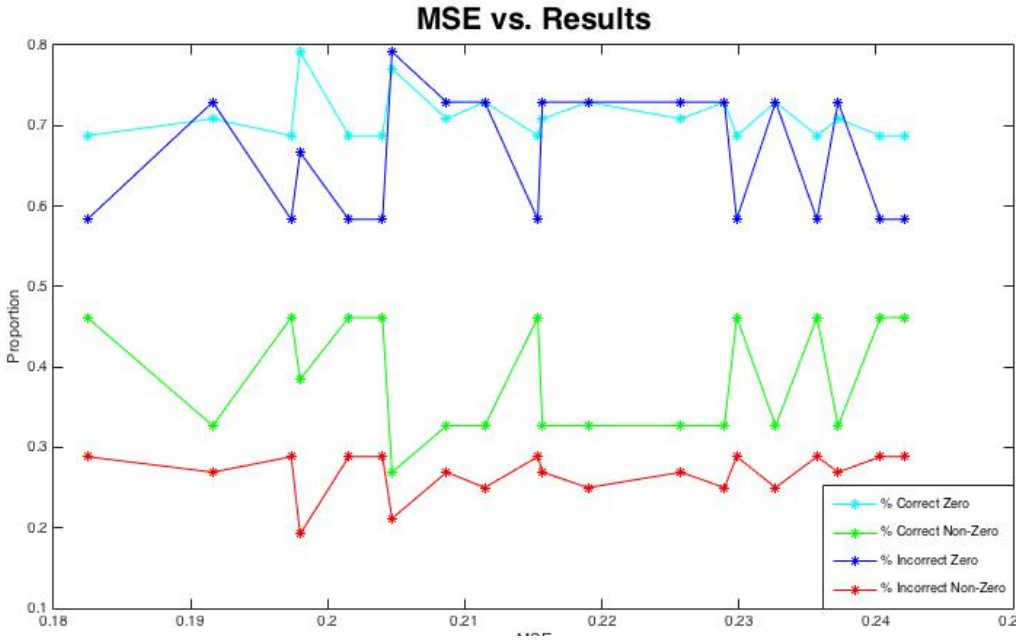
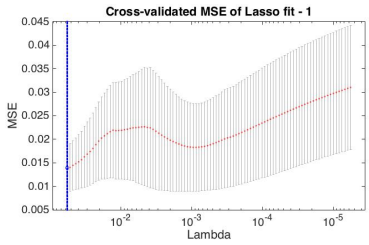
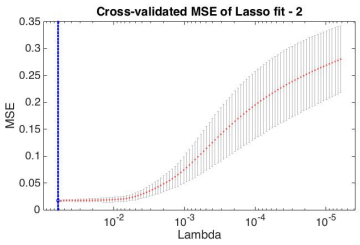


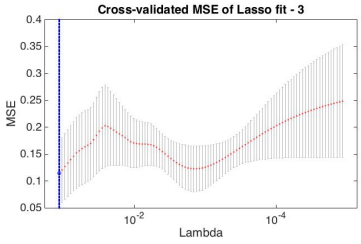
Figure 3: Results vs. MSE for MSE $\leqslant 0.25$



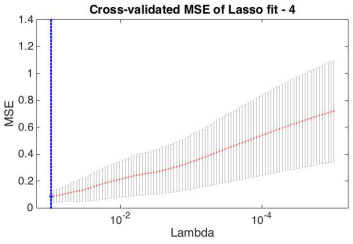
(a) Individual 1



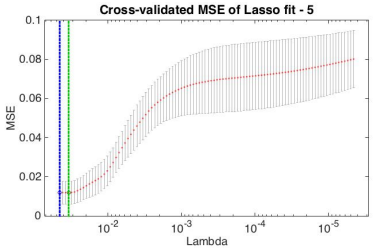
(b) Individual 2



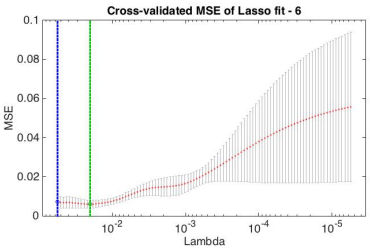
(c) Individual 3



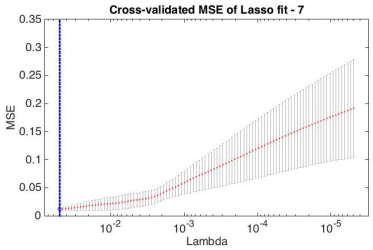
(d) Individual 4



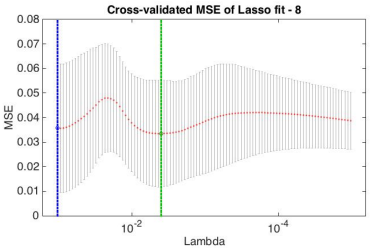
(e) Individual 5



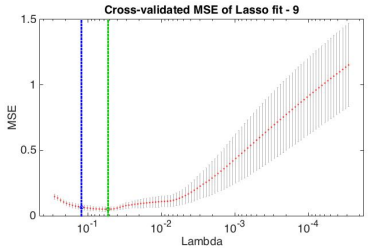
(f) Individual 6



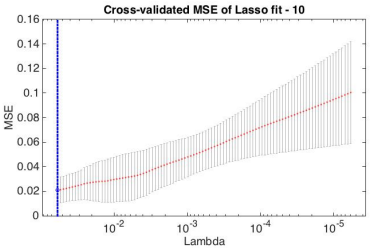
(g) Individual 7



(h) Individual 8



(i) Individual 9



(j) Individual 10

Figure 4: 5-Fold Cross-Validated paths for each Lasso computation for the best estimator.

8 Literature Review: Regression and the Lasso

The Lasso problem can be written as follows:

$$\begin{aligned} &\{1\} \quad \widehat{\beta_1} = \arg\min_{\beta}\{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\} \quad \text{or} \\ &\{2\} \quad \widehat{\beta_2} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2\} \quad \text{subject to} \quad \sum_j |\beta_j| \leqslant t \end{aligned} \tag{20}$$

where $\lambda, t \geqslant 0$ control the amount of "shrinkage" and are chosen through cross-validation.

Lemma. $\widehat{\beta_1} = \widehat{\beta_2}$

Proof. Problem {2} is typically referred to as the "primal" problem. In general, the primal problem is the constrained optimization problem of the form:

$$\widehat{\beta_2} = \arg\min f_0(\beta) \qquad \text{subject to} \quad f_1(\beta) \leqslant 0 \qquad \{f_{0,1} : \mathbb{R}^p \to \mathbb{R}\}$$

For us, $f_0(\beta) = \{\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2\}$ and $f_1(\beta) = \sum_j |\beta_j| - t$.

Problem {1} is typically referred to as the "dual" function problem. First, we define the Lagrangian function as:

$$L(\beta, \lambda) = f_0(\beta) + \lambda f_1(\beta) = \|Y - X\beta\|_2^2 + \lambda(\|\beta\|_1 - t) = \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 - \lambda t$$

The dual function is then defined as: $D(\lambda) = \min_\beta L(\beta, \lambda)$, $\{D : \mathbb{R} \to \mathbb{R}^p\}$.

Remark 1. Notice that $D(\lambda)$ is a minimization over β. Hence, subtracting a constant λt does not change the optimal solution. Hence:

$$\min_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 - \lambda t = \min_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \equiv D(\lambda) = (1)$$

Remark 2. Notice that since $f_1(\beta) \leqslant 0 \to L(\beta, \lambda) \leqslant f_0(\beta)$ for all feasible β. Then, $\min_\beta L(\beta, \lambda) \leqslant \min_\beta f_0(\beta) \to D(\lambda) \leqslant \widehat{\beta_2}$, by definition. That is, we have obtained a lower bound on the optimal solution.

Remark 3. We can then find the greatest lower bound on the optimal solution by, i.e. $\widehat{\beta_1} =$

$\max_\lambda D(\lambda) \rightarrow \widehat{\beta_1} \leqslant \widehat{\beta_2}$. The difference between the two optimal solutions is called the duality gap.

Remark 4. Slater's Condition: Given the formulation of the problem as in the beginning of the proof, if (a) $f_0(\beta)$ and $f_1(\beta)$ are convex and (b) there exists $\beta^* \in relint(dom(f_1))$ s.t. $f_1(\beta^*) < 0$, then the duality gap is zero, i.e. $\widehat{\beta_1} = \widehat{\beta_2}$.

Note that all L_s norms, $1 \leqslant s < \infty$ are convex. Hence, (a) is satisfied. Recall that $f_1(\beta) = \sum_j |\beta_j| - t$. Since t = 0 produces the trivial solution, $\widehat{\beta_2} = \mathbb{0}$, let $t > 0$. Then, $f_1(\mathbb{0}) = 0 - t < 0$. Therefore, Slater's Condition is satisfied.

□

The idea of "penalized regression" using a norm is not limited to the L_1 norm. Meinshausen et. al [13] describes some of the other regression estimates that have been proposed using the L_s norm, where s is typically in the range [0, 2]. A value of s = 2 leads to the ridge estimate. For $s \leqslant 1$, the estimates provide sparse solutions, while the optimization problem in (1) is only convex for $s \geqslant 1$. The s = 1 case is, therefore, the only value of s for which subset selection takes place while the optimization problem is still convex and hence feasible for high dimensional problems.

It must be noted that bounding by the L_1 norm in order to achieve sparsity is not restricted to the regression context. Rasmussen et al. [16] describes a wide range of models in the fields of biostatistics and computational mathematics where minimization of least squares is extended with this norm constraint, like sparse principal component analysis, sparse partial least squares, sparse canonical variate analysis and sparse linear discriminant analysis.

8.1 Sparsity Illustration

The Lasso estimate under orthonormal design can be shown to take the following form:

$$\hat{\beta_j} = sign(\beta_j^{OLS})(|\beta_j^{OLS}| - \lambda)^+ \tag{21}$$

where $(f)^+ = \max(0,f)$ and λ is determined by the L_1 condition. Hence, the lasso retains only the largest coefficients, discarding the "small" ones. Another way of seeing why the lasso produces zero coefficients is to consider the case with just two coefficients. The constraint region (i.e. $|x_1| + |x_2| \leqslant t$) is a rotated square in 2-D while residual squared error is a quadratic function with elliptical contours centered at the OLS estimates. The minimum of the sum of these two will occur when they intersect. With proper "shrinkage", this happens at a corner (β_1 or β_2 is zero), as seen in the figure below.

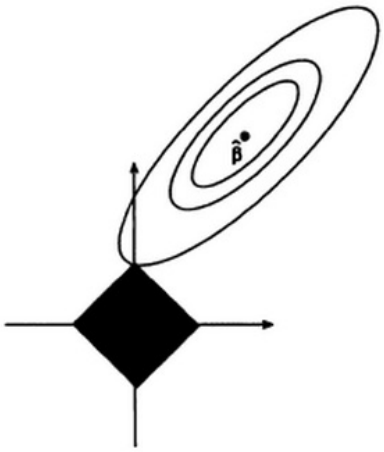


Figure 5: Constraint region and OLS contours for p = 2, from Tibshirani (1996).

8.2 Applications of the Lasso

The Lasso and its variants have been applied to many statistical problems in the last decade, most notably in computational biology and economics. Buhlmann and van de Geer [1], and the references therein, illustrate how the Lasso is used to predict DNA splice sites using a binary logistic regression and how it is used to identify the genes responsible for certain protein synthesis, among many others.

The Lasso is also found in the high-dimensional problems of macroeconomic research. For example, Li and Chen [12] apply the Lasso to forecasting macroeconomic variables and show that the LASSO-based dynamic factor models can reduce forecasting error. Specifically, they show that the group-LASSO outperforms simple dynamic factor models in out-of-sample forecast evaluations and reduces the complexity of these models.

8.3 High-Dimensional Lasso

Now I want to consider the "large p, small n", high-dimensional framework. In most model selection problems, it makes intuitive sense to allow the number of parameters to grow with sample size. Like Fan and Li [4], Fan and Peng [5] establish similar asymptotic properties of the nonconcave penalized likelihood methods as $p \to \infty$. Wang et al. [25] extend these results to the group Lasso when the variables are naturally grouped. Wei and Huang (2010) prove similar theorems with the adaptive Lasso while van de Geer [23] does so for generalized linear models.

Likewise, Meinshausen and Buhlmann [13] showed a similar result in the context of neigh-

borhood selection in Gaussian graphical models. Under a neighborhood stability condition on the design matrix and certain additional regularity conditions, they proved that the Lasso is consistent, even when the number of variables tends to infinity at a rate faster than n.

Zhao and Yu [29] formalized the neighborhood stability condition in the context of linear regression models as a "strong irrepresentable condition". They showed that under this crucial condition and certain other regularity conditions, the Lasso is consistent for variable selection, even when the number of variables p is as large as $\exp(n^a)$ for some $0 \leqslant a \leqslant 1$. This condition depends mainly on the covariance of the covariates. Namely, define

$$C^n = \begin{Bmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{Bmatrix}$$

where $C_{12}^n = \frac{1}{n} X_n(1)' X_n(2)$ represents the covariance between the p chosen covariates (X(1)) and the $p - q$ zero covariates (X(2)).

The strong irrepresentable condition states: there exists a positive constant vector ν

$$|C_{21}^n (C_{11}^n)^{-1} sign(\beta_{(1)}^n)| \leqslant \mathbf{1} - \nu$$

where $\mathbf{1}$ is a $p - q$ vector of 1's and the inequality holds element wise.

Zhang and Huang [28] provide a different set of sufficient conditions under which the Lasso is "rate consistent" in sparsity and bias of the selected model. They consider a model "sparse" if most coefficients are small, in the sense that the sum of their absolute values is below a certain level. This is slightly different than the usual definition of sparsity, where certain coefficients must be set exactly to zero. However, this framework is more general.

Define $\hat{q}$, $\hat{B}$, ζ_α, and B as the number of non-zero coefficients, the bias, a measure of large coefficients for the missing variables and a measure of the true missing variables, respectively.

$$\hat{q} = \#(j : \hat{\beta} \neq 0) \qquad \tilde{B} = \|(I - P)X\beta\|_2 \quad \zeta_\alpha = \left(\sum |\beta_j|^\alpha I(\hat{\beta} = 0) \right)^{1/\alpha}$$

The authors prove that

$$\hat{q} = O(q) \qquad \tilde{B} = O_p(B) \qquad \sqrt{n}\zeta_\alpha = O(B)$$

8.4 Estimating Lasso with dependent covariates

Up to this point, I have only considered the Lasso under an iid assumption. How does the Lasso perform when the covariates are dependent?

Gupta [6] extends the results of Fan and Li to regression models with a general weak dependence structure. He determines that the asymptotic distribution of the Lasso when p is fixed and the number of observations converges to infinity is a multivariate normal distribution, under an appropriate choice of the tuning parameter.

Under certain restrictions on the rate of increase of the covariates as well as the rate of increase of p, the author obtains finite sample error bounds. More, he obtains sign consistency of the Lasso even when p grows exponentially with n. Lastly, he provides the consistency and $n^{(1/2-d)}$ consistency of the Lasso in the case where p is fixed and is less than n, under certain assumptions on the covariates.

Hebiri and Lederer [8] show that correlations among the covariates strongly influence the optimal tuning parameters. They also show theoretically and through simulations that, for suitably chosen tuning parameters, the Lasso predicts well regardless of the level of correlation. Specifically, the higher the correlations are, the smaller the optimal tuning parameter is, which may influence the way practitioners cross-validate for the tuning parameter.

8.5 Estimating AR Models with the Lasso

This section details the existing literature on the Lasso in the time series setting of estimating the order and parameters of autoregressive (AR) models. Typically, both estimation and model fitting rely on the assumption of fixed dimensional parameters. That is, it is assumed that the order of the AR process is known beforehand and that a model selection procedure that sequentially fits models of increasing dimension is adequate. Information criteria, such as the BIC and the Aikaike, are typically used to compare these models. The limitations that these assumptions pose can be addressed by the Lasso, as the Lasso simultaneously chooses the order by setting some parameters to zero and estimates the non-zero ones.

Nardi and Rinaldo [14] consider the Lasso for autoregressive models when the number of parameters and the maximal possible lag grows with the sample size.

Their setup is the following. Let X_t be defined as the AR(p) process

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + Z_t \qquad t = 1, ..., n$$

Let $y = (X_1, ..., X_n)$, $\phi = (\phi_1, ..., \phi_p)$, and $Z = (Z_1, ..., Z_n)$. Define the n x p matrix X with entry X_{t-j} in the t^{th} row and j^{th} column, for t = 1, . . . , n and j = 1, . . . , p. The Lasso-type estimator $\hat{\phi_n}$ is defined to be the minimizer of:

$$\frac{1}{2n}||y - X\phi||^2 + \lambda_n \sum_{j=1}^{p} \lambda_{n,j}|\phi_j|$$

where $\lambda_{n,j}, j = 1, ..., p$ are specific tuning parameters associated with the predictors X(t-j). The authors show that the Lasso possesses model selection, estimation and prediction consistency under suitable assumptions.

Wang et al. [24] considers, under the fixed p scenario, the classic linear regression model with AR(p) errors:

$$y_t = x_t\beta + \epsilon_t \qquad \epsilon_t = \phi_1\epsilon_{t-1} + ... + \phi_q\epsilon_{t-q} + e$$

with the Lasso as the minimizer of:

$$Q_n(\beta, \phi) = \sum \left\{ y_t - x_t'\beta - \sum_{j=1}^{q} \phi_i\epsilon_{t-j} \right\}^2 + \lambda \sum_{j=1}^{p} |\beta_j| + \gamma \sum_{j=1}^{q} |\phi_j|$$

They propose an iterative profiling procedure for estimating this model, where Q_n is divided into two objective functions, each with one parameter vector unknown and one fixed.

Schmidt and Makalic [18] approached the AR modeling problem with a Bayesian view. They exploit the fact that sum-of-absolutes penalty implied by the LASSO is equivalent to using a Laplace distribution as a prior distribution over the parameters. The authors parameterize the Lasso in terms of partial autocorrelations and control for stationarity explicitly. They show that the Lasso performs well in terms of prediction accuracy when compared to the standard selection techniques.

8.6 Lasso and VAR

Consider the k-dimensional time series $y_t = (y_{1t}, y_{2t}, ..., y_{kt})$ with $t \in (1, n)$. A vector autoregressive model of order p, Var(p), is defined as:

$$y_t = \nu + \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + u$$

where each Φ is a k x k coefficient matrix, ν is k x 1 intercept vector and u is a white noise process.

Hsu et al. [9] propose multiple hybrid estimation strategies, which includes traditional infor-

mation criteria-driven selection along with the Lasso. For example, they suggest the use of AIC to select the best order for the VAR (in which multiple series are considered simultaneously) and estimate the coefficients using the Lasso. In addition, the authors propose a "top-down" and "bottom-up" subset selection procedures. The "top-down" method starts with a full model and sequentially attempts to reduce the order without reducing a certain information criterion. The "bottom-up" method does the opposite, where the model sequentially increases the order until the criterion stops increasing. The authors conduct a simulation study and find that the hybrid Lasso method, where the model is reduced by the AIC first, performs best.

Ren and Zhang [17] use the adaptive Lasso to select the order and estimate the coefficients of a VAR(p) model. Song and Bickel [20] propose the application of a group Lasso penalty in the context of large vector autoregressions, where regression coefficients in the same group are shrunk to zero jointly. Their method is able to do variable selection and lag selection simultaneously, and is robust to the initial choice of lags.

References

[1] Buhlmann, Peter and Sara van de Geer. "Statistics for High-Dimensional Data." Springer Series in Statistics. 2011.

[2] Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. "Least Angle Regression." The Annals of Statistics 32, no. 2 (April 1, 2004).

[3] Elhorst, J. "Specification and estimation of spatial panel data models." International Regional Science Review 26(3), 244–268. (2003)

[4] Fan, Jianqing, and Runze Li. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." Journal of the American Statistical Association 96, no. 456 (December 1, 2001).

[5] Fan, Jianqing, and Heng Peng. "Nonconcave Penalized Likelihood with a Diverging Number of Parameters." The Annals of Statistics 32, no. 3 (June 1, 2004).

[6] Gupta, Shuva. "A Note on the Asymptotic Distribution of LASSO Estimator for Correlated Data." Sankhya A 74, no. 1 (November 17, 2012).

[7] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. "The Elements of Statistical Learning." Springer Series in Statistics. 2008.

[8] Hebiri, Mohamed, and Johannes Lederer. "How Correlations Influence Lasso Prediction." IEEE Transactions of Information Theory 59, no. 3, (March, 2013).

[9] Hsu, Nan-Jung, Hung-Lin Hung, and Ya-Mei Chang. "Subset Selection for Vector Autoregressive Processes Using Lasso." Computational Statistics & Data Analysis 52, no. 7 (March 15, 2008).

[10] Kaul, Abhishek. "Lasso with Long Memory Regression Errors." Journal of Statistical Planning and Inference 153 (October 2014).

[11] Knight, Keith, and Wenjiang Fu. "Asymptotics for Lasso-Type Estimators." The Annals of Statistics 28, no. 5 (October 2000).

[12] Li, Jiahan, and Weiye Chen. "Forecasting Macroeconomic Time Series: LASSO-Based Approaches and Their Forecast Combinations with Dynamic Factor Models." International Journal of Forecasting 30, no. 4 (October 2014).

[13] Meinshausen, Nicolai, and Peter Buhlmann. "High-Dimensional Graphs and Variable Selection with the Lasso." The Annals of Statistics 34, no. 3 (June 1, 2006).

[14] Nardi, Y., and A. Rinaldo. "Autoregressive Process Modeling via the Lasso Procedure." Journal of Multivariate Analysis 102, no. 3 (March 2011).

[15] Qin et. al, 2013. Qin, Zhiwei, Katya Scheinberg, and Donald Goldfarb. "Efficient Block-Coordinate Descent Algorithms for the Group Lasso." Mathematical Programming Computation 5, no. 2 (March 31, 2013).

[16] Rasmussen, Morten Arendt, and Rasmus Bro. "A Tutorial on the Lasso Approach to Sparse Modeling." Chemometrics and Intelligent Laboratory Systems 119 (October 1, 2012).

[17] Ren, Yunwen, and Xinsheng Zhang. "Model Selection for Vector Autoregressive Processes via Adaptive Lasso." Communications in Statistics: Theory & Methods 42, no. 13 (August 15, 2013).

[18] Schmidt, Daniel F., and Enes Makalic. "Estimation of Stationary Autoregressive Models with the Bayesian LASSO." Journal of Time Series Analysis 34, no. 5 (September 1, 2013).

[19] Simon, Noah. "A sparse-group Lasso". Journal of computational and graphical statistics, 22 (2), p. 231.

[20] Song, Song, and Peter J. Bickel. "Large Vector Auto Regressions." arXiv:1106.3915 [q-Fin, Stat], June 20, 2011.

[21] Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological) 58, no. 1 (January 1, 1996).

[22] Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. 109(3), 475–494 (2001)

[23] van de Geer, Sara A. "High-Dimensional Generalized Linear Models and the Lasso." The Annals of Statistics 36, no. 2 (April 2008).

[24] Wang, Hansheng, Guodong Li, and Chih-Ling Tsai. "Regression Coefficient and Autoregressive Order Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69, no. 1 (February 1, 2007).

[25] Wang, Lichun, Yuan You, and Heng Lian. "Convergence and Sparsity of Lasso and Group Lasso in High-Dimensional Generalized Linear Models." Statistical Papers 56, no. 3 (July 3, 2014).

[26] Wei, Fengrong, and Jian Huang. "Consistent Group Selection in High-Dimensional Linear Regression." Bernoulli 16, no. 4 (November 2010).

[27] Yuan, Ming, and Yi Lin. "Model Selection and Estimation in Regression with Grouped Variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, no. 1 (February 1, 2006): 49–67.

[28] Zhang, Cun-Hui, and Jian Huang. "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression." The Annals of Statistics 36, no. 4 (August 2008).

[29] Zhao, Peng, and Bin Yu. "On Model Selection Consistency of Lasso." J. Mach. Learn. Res. 7 (December 2006).